Compositional Concept Representation Using SNOMED: Towards Further Convergence of Clinical Terminologies

Kent A. Spackman, M.D., Ph.D.[1,2], Keith E. Campbell, M.D., Ph.D.[3],
[1]College of American Pathologists, Northfield, IL;
[2]Oregon Health Sciences University, Portland, OR;
[3]Kaiser-Permanente, Oakland, CA

This paper describes several approaches to the expression and coding of clinical concepts as composites of elementary entities, and describes an approach based on SNOMED RT that may permit further convergence of clinical terminology efforts. We explain the shortcomings of previous approaches to compositional concept representation, as well as the reasons for SNOMED's current approach, which adopts a foundation based in description logics (DLs). The DL model has many advantages: it establishes a formal semantics for SNOMED assertions and suggests a syntax; it provides a basis for understanding expressiveness and computational complexity, through correspondence with known results from DLs; and it helps to clarify the relationships among existing concept representation methods in SNOMED, NHS Clinical Terms (formerly the Read Codes), and GALEN, making a path to convergence more clear.

INTRODUCTION

One of the main reasons to create a comprehensive clinical terminology is to facilitate the accurate representation of clinical detail, allowing accurate storage, retrieval and analysis of patient data. Natural descriptions of clinical details are richly varied, and it would be practically impossible to enumerate them all. It has long been recognized that it would be desirable to have a model that allows composition of clinical concepts from atomic elements [1,12]. SNOMED [2] has always allowed compositional encoding, but many authors have recognized shortcomings in SNOMED's compositional concept representation. For example, some have called for a syntax, and others have noted that it is possible to represent the same concept with more than one unique combination of codes. [3] We acknowledge that these are legitimate concerns, and show that these are just two of several problems with compositional models.

COMPOSITIONAL CONCEPT MODELS

Some authors speak of SNOMED concept composition as though it is a single well-defined model. However, in examining the literature on SNOMED and looking at different interpretations and implementations, we have identified at least three major variations in how compositional concepts have been represented in the past. In order to more fully understand these approaches to compositional concept representation using SNOMED, we name and describe them, along with their strengths and weaknesses, and also describe our current approach based on SNOMED RT [4]. In this paper, the four approaches are called Compositional Concept Models (CCM) 1 through 4:
CCM-1: Unconstrained composition
CCM-2: Multi-axial composition
CCM-3: Attribute-value composition
CCM-4: Foundational model composition

CCM-1: Unconstrained Composition

CCM-1 might be called "unconstrained" concept composition. The basic idea is that elementary or atomic concepts are enumerated and classified in a nomenclature, and then a compositional concept can be constructed by combining more than one atomic concept. Even though the atomic concepts may be from different "axes," in CCM-1 there are no significant constraints on how the combination is to take place; the structure is simple concatenation. Interpretation of the meaning of the concatenated string usually is dependent on the knowledge of the individual who examines the string; computer-based interpretation of such compositional concepts is fraught with ambiguities and duplications. CCM-1 has been criticized by many authors. Two of the main criticisms are: 1) that a given concept can be represented many different ways, and 2) that it is not possible for the computer to recognize the equivalance of these different ways of representing the concept.

1091—8280/98/$5.00

Many of the studies of SNOMED's expressiveness seem to have assumed an unconstrained (CCM-1) compositional model [2]; it is possible, but has not been determined whether a more constrained or principled model of composition would have resulted in less expressiveness in these studies.

CCM-2: Multi-axial Composition

CCM-2 might be called "multi-axial" composition. This model was described in detail in the SNOMED II Coding Manual of 1979 [5]. The essence of the model is that there is a set of "axes" that can be combined to form composite concept representations or assertions. CCM-2 and CCM-1 are not often differentiated; however, CCM-2 is much stronger semantically, and also has less expressive flexibility.

This model has a long history in coding systems. The original SNOMED, published in 1976-77, was based in part on SNOP (1965) and SNDO, both of which had a "multi-axial" nature. For example, anatomic site (T for topography) and structural change (M for morphology) are separately enumerated (in SNOP, ICD-O, and SNOMED), and their basic elements can be combined. Thus, instead of having a separate code for every possible tumor morphology in every possible anatomic location, one simply combines the morphology with the topography. For example, adenocarcinoma of the stomach would be coded as a combination of the M-code for adenocarcinoma with the T-code for stomach.

The axes in SNOMED II are Procedure, Topography, Morphology, Etiology, Function, and Disease. Each axis is given a single field in the coding table, and three additional fields are added to the table: the "context," represented by Information Qualifiers (IQ), the time (duration), and finally linkages to other concepts. Each individual assertion or concept is represented as a row in a table, and combined assertions can be represented by linking successive rows together.

Figure 1a shows the SNOMED II coding template, and Figure 1b shows the representation of a composite concept using this template.

IQ	P	T	M	E	F	D	TIME	LINK

Figure 1a: SNOMED II Coding Template

IQ	P	T	M	E	F	D	TIME	LINK
(FD) FINAL DIAGNOSIS		T-64300 DUODENUM	M-34000 OBSTRUCTION, NOS					(DT) DUE TO
(FD) FINAL DIAGNOSIS		T-58700 AMPULLA OF VATER	M-81403 ADENO-CARCINOMA					

Figure 1b: SNOMED II Coding Template showing the codes for a final diagnosis of duodenal obstruction due to adenocarcinoma of the ampulla of Vater.

CCM-3: Attribute-value Composition

CCM-3 might be called "attribute-value" composition. The need for explicit attributes, instead of simply a list of a few axes, was apparent to the authors of the SNOMED II coding manual. Where CCM-2 conflated the axis and the attribute, CCM-3 splits them out, as in the example in figures 2a and 2b, which represent the concept "gunshot wound of forehead, by handgun, with hypovolemic shock, homicide."

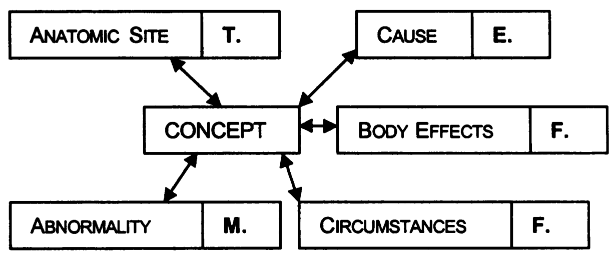

Figure 2a: Basic multi-axial concept template (from SNOMED II Coding Manual, Figure 6)

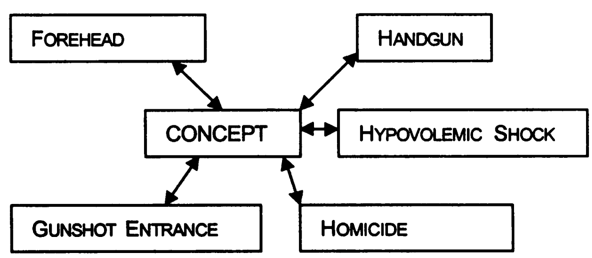

Figure 2b: Template from 2a with specific examples (from SNOMED II Coding Manual, Figure 8)

It is clear from these two figures that we have an example of a single concept that requires two codes

from the same axis (F). Note that such a representation cannot properly be made using the SNOMED II Coding Template (CCM-2). Figure 2a explicitly lists the attributes, while figure 2b lists the values, and thus the concept may be represented as:
Concept =

Anatomic site: Forehead
Abnormality: Gunshot entrance
Cause: Handgun
Circumstances: Homicide
Body Effects: Hypovolemic shock

Examining this representation, it is clear that each line has two elements, the first of which describes an attribute and the second of which describes the value of that attribute. In the terminology of description logic, the first element may be called the "defining relationship" and the second element may be called the "value restriction.." [6] In the example, "circumstances" could be regarded as a defining relationship, and "homicide" a value restriction.

One way of interpreting CCM-2 that helps to understand the origin of its basic weakness is that it enumerates and classifies the values, but leaves the attributes implicit (in the axis) or undefined. In order for there to be clear and unambiguous representation of concepts, and consistent use of the terminology, each compositional concept expression should explicitly give both the attributes and their values. This is what CCM-3 does that CCM-2 does not do.

The total number of possible attributes is much larger than the number of "axes," or orthogonal semantic groupings of basic concepts. Conflating the attributes with the axes results in limited expressiveness. Thus CCM-3 has more flexibility and expressiveness than CCM-2, and also has a stronger explicit semantics. CCM-3 corresponds quite closely to the NHS Clinical Terms compositional model, as described in the "Version 3.1 File Structure: Qualifier Extensions".

The SNOMED III G axis (general linkage and modifiers) contains a relatively large collection of terms that might be used as "attributes" in an attribute-value type of compositional concept representation. However, particular concepts are not explicitly identified as attributes, and no coherent effort was made to identify a set of such attributes (or defining relationships) that, taken together, could form the basis for a "foundational model" of composite concept representation. Thus CCM-3 lacks a consistent set of defining relationships as part of the compositional model. CCM-4 addresses this deficiency.

CCM-4: Foundational Models, Description Logics

CCM-4 may be called "foundational model" composition. [1] It builds upon CCM-3 by adding a formal semantics based on description logics, and by explicitly requiring a foundational set of defining relationships. The difference between CCM-3 and CCM-4 is therefore primarily not in the format of the representation of composite concepts, but in the models underlying the format.

Foundational models

Foundational models consist of sets of defining relationships, along with the assumptions and conceptual model that underlies these relationships. For example, the SNOMED multi-axial coding of tumors can be re-cast into CCM-4 by declaring the defining relationships "has-topography" and "has-morphology" as the core of the foundational model for compositional encoding of tumors.

Description logics

Description logics are formal subsets of predicate logic which typically have a formal semantics based on Tarski-style denotational semantics. [7] A more comprehensive review of the characteristics of description logics is beyond the scope of this paper, and can be found elsewhere. [8] However, a few basic definitions should suffice to describe the importance of description logic as a foundation for a compositional model of medical concepts.

Description logic statements are used to denote the essential characteristics of concepts; that is, to create a formal representation of the semantic definition of a concept, based on those features or characteristics that are always true and that differentiate one concept from another.

Description logic statements can be composed of "concept-forming operators," that is, operators that take concepts and defining relationships (also known as "roles") and combine them to form new logical expressions that define the meaning of a concept.

A description logic "engine" reads DL statements and then computes subsumption relationships between concepts; in other words, the DL engine can tell from the DL definition of two terms whether one is a specialization or generalization of the other.

It has been shown that computing subsumption for expressive DLs is computationally intractable; in order to achieve tractable (worst-case) subsumption, the concept-forming operators used in the DL must be restricted.

Horrocks describes an experiment in which he attempts to compare the GRAIL language with LOOM, another description logic language, and he describes the concept-forming operators that are used. [7] CCM-4 is based on a tractable description logic which uses the same set of concept-forming operators (K-REP). [9] In Table 1, commonly-applied concept-forming operators are listed, and the concept-forming operators used by CCM-4 and by GRAIL are listed with an asterisk.

	Operator name	Notation
1*	Top (everything)	$\top$
2*	Bottom (∅)	$\perp$
3*	Conjunction	$C_1 \sqcap ... \sqcap C_n$
4*	Exists restriction	$\exists R.C$
5.	All restriction	$\forall R.C$
6.	Disjunction	$C_1 \sqcup ... \sqcup C_n$
7.	Negation	$\neg C$
8.	Number restriction	$\geq nR.C$
9.	Number restriction	$\leq nR.C$

Table 1. General concept-forming operators in description logics. *=operators used in GRAIL and in CCM-4.

We can now show how to solve the "acute appendicitis" example, commonly used as an example of the inadequacies of CCM-1 and CCM-2. Table 2 shows the relevant SNOMED codes and terms.

D5-46210	Acute appendicitis, NOS
D5-46100	Appendicitis, NOS
G-A231	Acute
M-40000	Inflammation, NOS
M-41000	Acute inflammation, NOS
T-59200	Appendix, NOS
G-C006	Has location (In)

Table 2: SNOMED Codes related to the concept "Acute appendicitis"

Tables 3 through 6 show the different possible representations of acute appendicitis using each of the compositional concept models described in this paper.

D5-46210	Acute appendicitis
G-A231,D5-46100	Acute + appendicitis
M-41000, G-C006, T-59200	Acute inflammation + In + appendix
G-A231, M-40000, G-C006, T-59200	Acute + inflammation + In + appendix

Table 3: Acute appendicitis represented by CCM-1

IQ	P	T	M	E	F	D	T	L
						D5-46210 Acute Appendicitis		
G-A231 Acute						D5-46100 Appendicitis		
		T-59200 Appendix	M-41000 Acute Inflammation					
G-A231 Acute		T-59200 Appendix	M-40000 Inflammation					

Table 4: Acute appendicitis represented by CCM-2

D5-46210		
D5-46100	has-course	G-A231
M-41000	assoc-topography	T-59200
M-40000	has-course	G-A231
	assoc-topography	T-59200

Table 5: Acute appendicitis represented by CCM-3

D5-46210		
D5-46100	has-course	G-A231
DF-00000	assoc-topography	T-59200
	assoc-morphology	M-41000
DF-00000	has-course	G-A231
	assoc-topography	T-59200
	assoc-morphology	M-40000

Table 6: Acute appendicitis represented by CCM-4

Note that because CCM-4 has a foundational model of disease that links all disease expressions to a "root" concept of disease (DF-00000), the last two CCM-3 representations are explicitly invalid as representations of disease.

TERMINOLOGY CONVERGENCE

NHS Clinical Terms version 3 adopts a object-attribute-value triple approach to representing concepts. [11] Thus it explicitly identifies attributes, which may be interpreted as "defining relationships." In fact, these are used as the basis for "semantic definitions."

The same tables used to express semantic definitions are also used to create templates for composition of concepts. Each term that can be modified is listed with the defining attributes that may modify it, and the values (or value sets, by reference) that may be used. Convergence would be possible based on harmonization of the attributes in these template files with the SNOMED foundational models.

GRAIL uses the same concept-forming operators as CCM-4. [7] This limited set of concept-forming operators seems to be satisfactory for a significant part of terminological representation in health and medicine. In addition, GRAIL has three kinds of sanctioning: 1) conceivable, 2) grammatical, and 3) sensible. [10] CCM-4 provides conceivable sanctioning through creation of defining roles and value restrictions. Sensible sanctioning is required primarily for generative terminology, which in turn is mainly a user-interface issue rather than a reference terminology issue.

Further examination of the GRAIL CORE model, and the intermediate representation being used for clinical modelers, may allow consideration of further convergence of it and SNOMED's foundational models.

CONCLUSION

Reliable and accurate compositional concept representation is now feasible through the combined use of a reference terminology, description logic semantics, and a set of foundational models. Implementation of this approach in the Kaiser-Permanente Convergent Medical Terminology project will afford ample opportunity to evaluate its effectiveness.

Acknowledgments

This work is supported in part by a grant from the College of American Pathologists, and in part by Kaiser Permanente.

References

1. Campbell KE, Das AK, Musen MA. A Logical foundation for representation of clinical data. J Am Med Informatics Assoc 1:218-232, 1994.

2. Cote RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L, eds. The Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International. Northfield, IL: College of American Pathologists, 1993.

3. Chute CG, et al. The content coverage of clinical classifications. JAMIA 3:224-233, 1996.

4. Spackman KA, Campbell KE, Cote RA. SNOMED RT: A reference terminology for health care. In: AMIA Annual Fall Symposium, 640-644, 1997.

5. Gantner GE, Côté RA, Beckett RS, eds. Systematized Nomenclature of Medicine, Coding Manual. Skokie, IL: College of American Pathologists, 1979.

6. Dolin RH, et al. Evaluation of a "lexically assign, logically refine" strategy for semi-automated integration of overlapping terminologies. JAMIA 5:203-213, 1998.

7. Horrocks IR. A comparison of two terminological knowledge representation systems. University of Manchester thesis. July 1995.

8. Woods WA, Schmolze JG. The KL-ONE family. Computers and Mathematics with Applications -- Special Issue on Artificial Intelligence, 23(2-5):133-177, 1992.

9. Mays E, Dionne R, Weida R. K-REP system overview. SIGART Bulletin 2(3):88-92, 1991.

10. Rector AL, Bechhofer S, Goble CA, et al. The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine* 9(2):139-171, 1997.

11. NHS Centre for Coding and Classification. Read Codes File Structure Version 3.1 - The Qualifier Extensions. January 1995.

12. Evans DA, et al. Toward a medical concept representation language. JAMIA 1:207, 1994.